Bayesian Hierarchical Models and the Maximum Entropy Principle

Brendon J. Brewer

The University of Auckland

MaxEnt

This is my favourite interpretation of maximum entropy, which I got from Ariel Caticha.

MaxEnt

This is my favourite interpretation of maximum entropy, which I got from Ariel Caticha.

Starting from a prior distribution $\pi(\boldsymbol{x})$, and given a constraint on probability distributions, the updated distribution is given by maximising the relative entropy

$$H(p;\pi) = -\sum_i p(\boldsymbol{x}) \log\left(\frac{p(\boldsymbol{x})}{\pi(\boldsymbol{x})}\right) \tag{1}$$

subject to the given constraint and the normalisation condition.

MaxEnt

If you select the right constraint — that the probability of the 'data' proposition should now be 1, this is equivalent to the prior $\rightarrow$ posterior update of Bayesian inference, but is more general.

MaxEnt

If you select the right constraint — that the probability of the 'data' proposition should now be 1, this is equivalent to the prior $\rightarrow$ posterior update of Bayesian inference, but is more general.

However, the generalisation is hard to use in practice because it is unclear where constraints that refer to probabilities (Jaynes' "testable information" which is actually untestable) would come from.

Expectation constraints

Consider some function $f(\boldsymbol{x})$ whose expected value is given (somehow). The updated distribution is given by the well known result:

$$p(\boldsymbol{x}) \propto \pi(\boldsymbol{x}) \exp\left[\lambda f(\boldsymbol{x})\right]. \tag{2}$$

The value of λ is selected to ensure that the correct expected value of $f()$ is obtained.

Expectation constraints: general version

With more than one expectation specified:

$$p(\boldsymbol{x}) \propto \pi(\boldsymbol{x}) \exp\left[\sum_i \lambda_i f_i(\boldsymbol{x})\right]. \tag{3}$$

The values of the λs ares selected to ensure that the correct expected values of $f_i()$ are obtained.

Exponential Example

Consider n positive quantities $x_1, ..., x_n$, and an expectation constraint on the average T:

$$\langle T \rangle = \left\langle \frac{1}{n} \sum_{i=1}^{n} x_i \right\rangle = \mu \tag{4}$$

Exponential Example

Consider n positive quantities $x_1, ..., x_n$, and an expectation constraint on the average T:

$$\langle T \rangle = \left\langle \frac{1}{n} \sum_{i=1}^{n} x_i \right\rangle = \mu \tag{4}$$

The MaxEnt result is independent exponential distributions:

$$p(\boldsymbol{x}) = \prod_{i=1}^{n} \frac{1}{\mu} \exp\left(-\frac{x_i}{\mu}\right). \tag{5}$$

Hierarchical Reflex

People often think "but what if μ is really unknown, so this is only the prior conditional on μ?". This leads to a hierarchical model with

$$p(\mu, \boldsymbol{x}) = p(\mu) \prod_{i=1}^{n} \frac{1}{\mu} \exp\left(-\frac{x_i}{\mu}\right), \tag{6}$$

but the MaxEnt interpretation is lost. Can we bring it back?

Hierarchical Reflex

People often think "but what if μ is really unknown, so this is only the prior conditional on μ?". This leads to a hierarchical model with

$$p(\mu, \boldsymbol{x}) = p(\mu) \prod_{i=1}^{n} \frac{1}{\mu} \exp\left(-\frac{x_i}{\mu}\right), \tag{6}$$

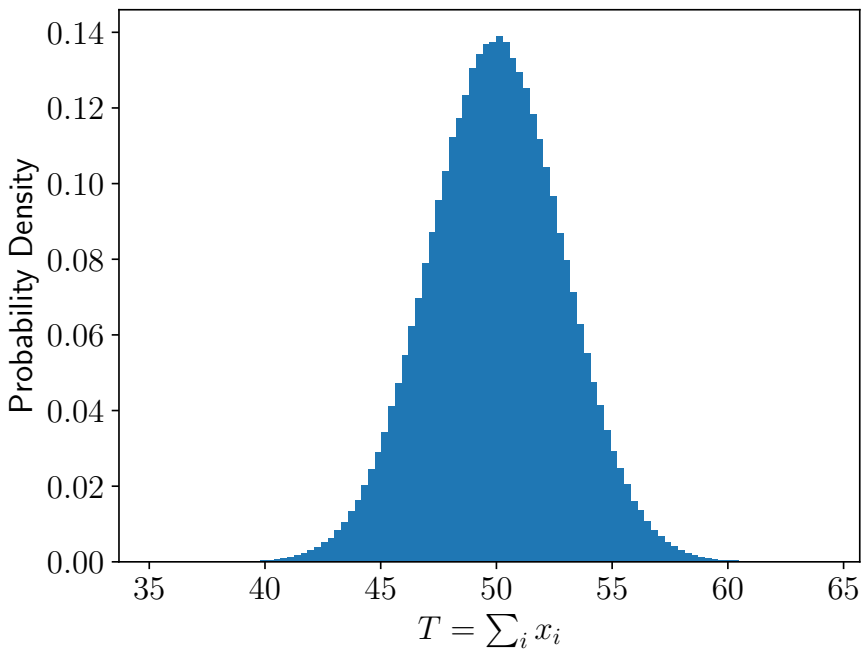
but the MaxEnt interpretation is lost. Can we bring it back?

Answer

Yes, provided we identify the right constraint.

A Different Type of Constraint

Consider a flat prior $\pi(\boldsymbol{x}) \propto 1$ over a wide domain. This implies a certain prior for the arithmetic mean $T = \frac{1}{n}\sum_i x_i$ that is too narrow and informative (by the central limit theorem). Example with $n = 100$ dimensions and Uniform$(0, 100)$ prior for each x_i:



Goal

What if we want to control the implied marginal distribution of T, but otherwise keep the distribution as close as possible to π?

Simple Example — Three-Valued Function

Starting with $\pi(\boldsymbol{x})$, suppose we want to control the marginal distribution of $f(\boldsymbol{x})$. For simplicity, assume $f(\boldsymbol{x})$ can only take values 1, 2, or 3.

Simple Example — Three-Valued Function

Starting with $\pi(\boldsymbol{x})$, suppose we want to control the marginal distribution of $f(\boldsymbol{x})$. For simplicity, assume $f(\boldsymbol{x})$ can only take values 1, 2, or 3. Under any distribution $p(\boldsymbol{x})$, the probability that $f(\boldsymbol{x}) = 1$ is

$$P(f(\boldsymbol{x}) = 1) = \int p(\boldsymbol{x}) \mathbb{1}\left(f(\boldsymbol{x}) = 1\right) \, d\boldsymbol{x} \tag{7}$$

Simple Example — Three-Valued Function

Starting with $\pi(\boldsymbol{x})$, suppose we want to control the marginal distribution of $f(\boldsymbol{x})$. For simplicity, assume $f(\boldsymbol{x})$ can only take values 1, 2, or 3. Under any distribution $p(\boldsymbol{x})$, the probability that $f(\boldsymbol{x}) = 1$ is

$$P(f(\boldsymbol{x}) = 1) = \int p(\boldsymbol{x}) \mathbb{1}\left(f(\boldsymbol{x}) = 1\right) \, d\boldsymbol{x} \tag{7}$$

Similarly for 2 and 3 we get

$$P(f(\boldsymbol{x}) = 2) = \int p(\boldsymbol{x}) \mathbb{1}\left(f(\boldsymbol{x}) = 2\right) \, d\boldsymbol{x} \tag{8}$$

$$P(f(\boldsymbol{x}) = 3) = \int p(\boldsymbol{x}) \mathbb{1}\left(f(\boldsymbol{x}) = 3\right) \, d\boldsymbol{x} \tag{9}$$

Simple Example — Three-Valued Function

Starting with $\pi(\boldsymbol{x})$, suppose we want to control the marginal distribution of $f(\boldsymbol{x})$. For simplicity, assume $f(\boldsymbol{x})$ can only take values 1, 2, or 3. Under any distribution $p(\boldsymbol{x})$, the probability that $f(\boldsymbol{x}) = 1$ is

$$P(f(\boldsymbol{x}) = 1) = \int p(\boldsymbol{x}) \mathbb{1}\left(f(\boldsymbol{x}) = 1\right) \, d\boldsymbol{x} \tag{7}$$

Similarly for 2 and 3 we get

$$P(f(\boldsymbol{x}) = 2) = \int p(\boldsymbol{x}) \mathbb{1}\left(f(\boldsymbol{x}) = 2\right) \, d\boldsymbol{x} \tag{8}$$

$$P(f(\boldsymbol{x}) = 3) = \int p(\boldsymbol{x}) \mathbb{1}\left(f(\boldsymbol{x}) = 3\right) \, d\boldsymbol{x} \tag{9}$$

These are just three expected values!

Simple Example — Three-Valued Function — Solution

To control the three probabilities (now cast as expected values), MaxEnt gives the solution

$$p(\boldsymbol{x}) \propto \pi(\boldsymbol{x}) \exp\left[\lambda_1 \mathbb{1}\left(f(\boldsymbol{x}) = 1\right) + \lambda_2 \mathbb{1}\left(f(\boldsymbol{x}) = 2\right) + \lambda_3 \mathbb{1}\left(f(\boldsymbol{x}) = 3\right)\right]. \tag{10}$$

We tweak λ_1, λ_2, and λ_3 to get the desired probabilities.

Simple Example — Three-Valued Function — Solution

To control the three probabilities (now cast as expected values), MaxEnt gives the solution

$$p(\boldsymbol{x}) \propto \pi(\boldsymbol{x}) \exp\left[\lambda_1 \mathbb{1}\left(f(\boldsymbol{x}) = 1\right) + \lambda_2 \mathbb{1}\left(f(\boldsymbol{x}) = 2\right) + \lambda_3 \mathbb{1}\left(f(\boldsymbol{x}) = 3\right)\right]. \tag{10}$$

We tweak λ_1, λ_2, and λ_3 to get the desired probabilities.
Note that the expression inside the exp is just a funny way of expressing a mapping from f-values to λ values (statisticians may recognise it from regression models involving a factor).

General Solution

From this example, it seems obvious to me (but I haven't proved it) that the general solution to this problem is

$$p(\boldsymbol{x}) \propto \pi(\boldsymbol{x}) \exp\left[g\left(f(\boldsymbol{x})\right)\right], \tag{11}$$

where the function $g()$ is chosen so that we get the desired marginal distribution for $f(\boldsymbol{x})$.

General Solution

From this example, it seems obvious to me (but I haven't proved it) that the general solution to this problem is

$$p(\boldsymbol{x}) \propto \pi(\boldsymbol{x}) \exp\left[g\left(f(\boldsymbol{x})\right)\right], \tag{11}$$

where the function $g()$ is chosen so that we get the desired marginal distribution for $f(\boldsymbol{x})$.
The exp can be absorbed into the function g if we want, giving a simpler result:

$$p(\boldsymbol{x}) \propto \pi(\boldsymbol{x}) g\left(f(\boldsymbol{x})\right). \tag{12}$$

Exponential Example

Returning to the exponential example with a hierarchical prior, the joint prior for the hyperparameter and the parameters is

$$p(\mu, \boldsymbol{x}) = p(\mu)p(\boldsymbol{x} \mid \mu) \tag{13}$$

$$= p(\mu)\mu^{-n} \exp\left(-\frac{\sum_i x_i}{\mu}\right) \tag{14}$$

Exponential Example

Returning to the exponential example with a hierarchical prior, the joint prior for the hyperparameter and the parameters is

$$p(\mu, \boldsymbol{x}) = p(\mu)p(\boldsymbol{x} \mid \mu) \tag{13}$$

$$= p(\mu)\mu^{-n} \exp\left(-\frac{\sum_i x_i}{\mu}\right) \tag{14}$$

and the marginal prior for $\boldsymbol{x}$ is

$$p(\boldsymbol{x}) = \int_0^\infty p(\mu)\mu^{-n} \exp\left(-\frac{\sum_i x_i}{\mu}\right) \, d\mu \tag{15}$$

Exponential Example

The marginal prior for $\boldsymbol{x}$ is

$$p(\boldsymbol{x}) = \int_0^\infty p(\mu)\mu^{-n} \exp\left(-\frac{\sum_i x_i}{\mu}\right) d\mu \tag{16}$$

$$p(\boldsymbol{x}) = \int_0^\infty p(\mu)\mu^{-n} \exp\left(-\frac{nT}{\mu}\right) d\mu \tag{17}$$

which only depends on $\boldsymbol{x}$ through $T = \frac{1}{n}\sum_i x_i$.

Exponential Example

The marginal prior for $\boldsymbol{x}$ is

$$p(\boldsymbol{x}) = \int_0^\infty p(\mu)\mu^{-n} \exp\left(-\frac{\sum_i x_i}{\mu}\right) d\mu \tag{16}$$

$$p(\boldsymbol{x}) = \int_0^\infty p(\mu)\mu^{-n} \exp\left(-\frac{nT}{\mu}\right) d\mu \tag{17}$$

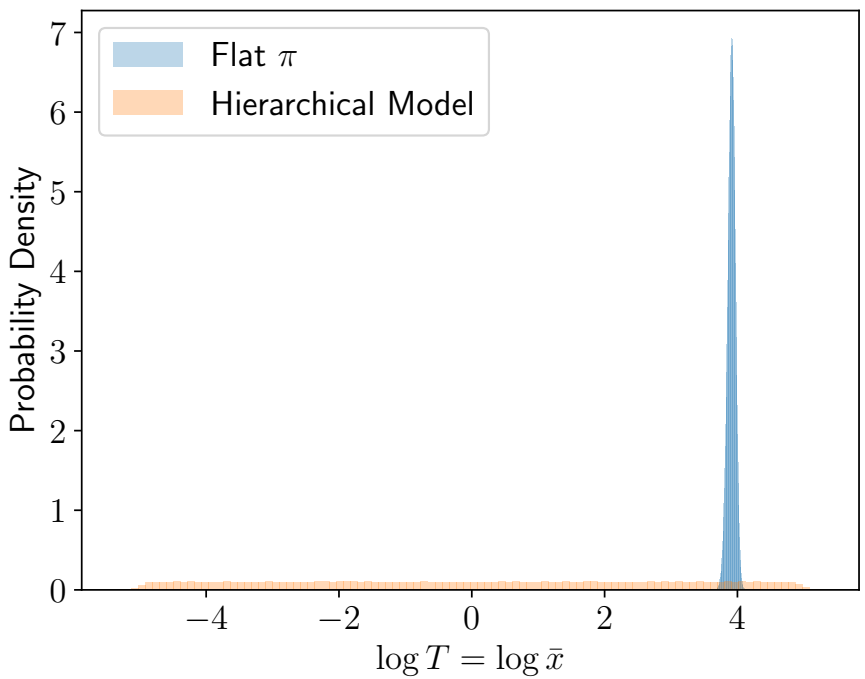
which only depends on $\boldsymbol{x}$ through $T = \frac{1}{n}\sum_i x_i$.This is therefore a distribution of the form

$$p(\boldsymbol{x}) \propto \pi(\boldsymbol{x}) g\left(\frac{1}{n}\sum_i x_i\right) \tag{18}$$

and is MaxEnt with uniform π and a specified marginal distribution for $T = \frac{1}{n}\sum_i x_i$.

What is the Implied Prior on T?

It is difficult to find the $g()$ function for a specified marginal prior, but if we instead just put $\ln(\mu) \sim \text{Uniform}(-5, 5)$ for example, we get something almost log-uniform for $T = \frac{1}{100}\sum_{i=1}^{100} x_i$:



Moral

Moral

Using the hierarchical model is equivalent to using MaxEnt with this specified marginal prior for $T = \frac{1}{n}\sum_i x_i$, fixing the problem with the prior implied by π.

Gaussian Example

Suppose $\pi(\boldsymbol{x}) \propto 1$ over a wide range. This implies inappropriately narrow priors for $T_1 = \sum_i x_i$ and $T_2 = \sum_i x_i^2$ which may be two quantities of interest.

Gaussian Example

Suppose $\pi(\boldsymbol{x}) \propto 1$ over a wide range. This implies inappropriately narrow priors for $T_1 = \sum_i x_i$ and $T_2 = \sum_i x_i^2$ which may be two quantities of interest.

Using an extension of the previous result, we can constrain the implied prior for T_1 and T_2 and obtain the MaxEnt distribution:

$$p(\boldsymbol{x}) \propto \pi(\boldsymbol{x}) g\left(f_1(\boldsymbol{x}), f_2(\boldsymbol{x})\right) \tag{19}$$

$$= \pi(\boldsymbol{x}) g\left(\sum_i x_i, \sum_i x_i^2\right). \tag{20}$$

Gaussian Example

If these were the usual expectation constraints we would have a gaussian distribution over $\boldsymbol{x}$, given a μ and a σ. If we let μ and σ be unknown (hierarchical) we get this marginal prior for the $\boldsymbol{x}$ quantities:

$$p(\boldsymbol{x}) = \int p(\mu, \sigma) \prod_i \text{Normal}(\boldsymbol{x}; \mu, \sigma^2)\, d\mu\, d\sigma \tag{21}$$

Gaussian Example

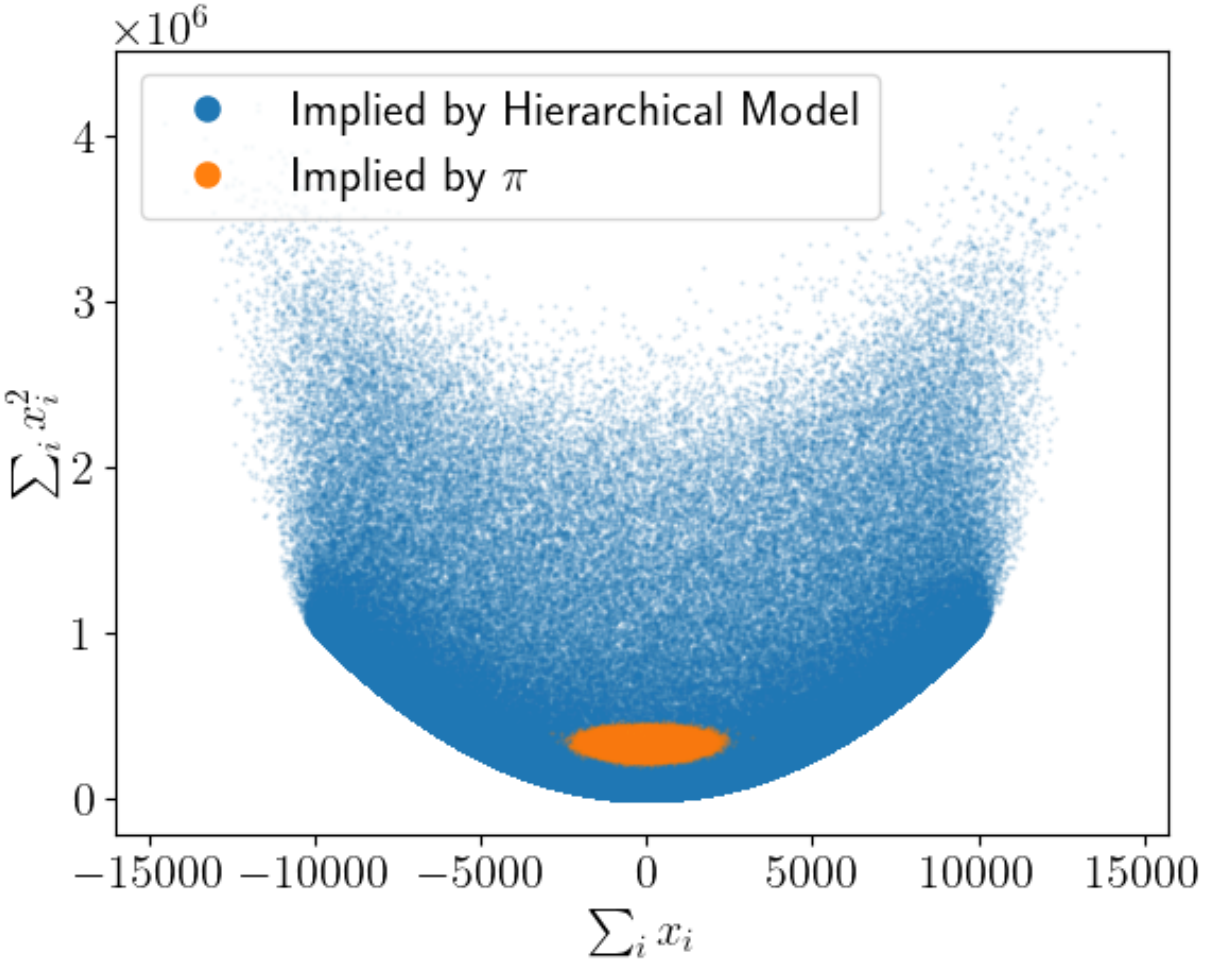
If these were the usual expectation constraints we would have a gaussian distribution over $\boldsymbol{x}$, given a μ and a σ. If we let μ and σ be unknown (hierarchical) we get this marginal prior for the $\boldsymbol{x}$ quantities:

$$p(\boldsymbol{x}) = \int p(\mu, \sigma) \prod_i \text{Normal}(\boldsymbol{x}; \mu, \sigma^2)\, d\mu\, d\sigma \tag{21}$$

This depends on $\boldsymbol{x}$ only through the 'sufficient statistics' T_1 and T_2 and is thus MaxEnt, of the form

$$p(\boldsymbol{x}) \propto \pi(\boldsymbol{x}) g\left(\sum_i x_i, \sum_i x_i^2\right). \tag{22}$$

The Two Priors



Moral

Moral

Using the hierarchical model is equivalent to using MaxEnt with this specified marginal prior for the sum and sum-of-squares of the x-values, fixing the problem with the prior implied by π.

Conclusions

- At least in some cases, hierarchical models can be thought of as MaxEnt distributions, incorporating a constraint on the marginal distribution of some function of the unknowns.

Conclusions

- At least in some cases, hierarchical models can be thought of as MaxEnt distributions, incorporating a constraint on the marginal distribution of some function of the unknowns.
- The MaxEnt solution for this case is simple with the prior multiplied by a transformed version of the function of interest.

Conclusions

- At least in some cases, hierarchical models can be thought of as MaxEnt distributions, incorporating a constraint on the marginal distribution of some function of the unknowns.
- The MaxEnt solution for this case is simple with the prior multiplied by a transformed version of the function of interest.
- This argument does not rely on any asymptotics, unlike exchangability/de Finetti justifications for hierarchical models.